Predicting ecosystem emergent properties at multiple scales

Jack A Gilbert[1,2,3,4] and Chris Henry[5,6]

[1]Institute for Genomic and Systems Biology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A.
[2]Department of Ecology and Evolution, University of Chicago, 1101 E 57th Street, Chicago, IL 60637
[3]Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA
[4]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, 310058, China
[5]Computation Institute, University of Chicago, 5735 S Ellis Ave, Chicago, IL 60637, U.S.A.
[6]Mathematics and Computer Science Division, 9700 S. Cass Ave, Argonne, IL 60439, U.S.A.

Biological phenomena at the microbial community level encode information about the subpopulation of cells and taxa at a specific time in the succession and biogeochemical evolution of that assemblage. To capture and understand the entire population for a community at a temporal resolution at which biogeochemical processes influence geological climate dynamics, requires large-scale computational simulations of their formation and evolution. The interactions between components of biology, geochemistry and physical processes within an ecosystem are inherently non-linear, with complex feedback mechanisms. However, this complexity does not preclude quantification of the dynamics that govern the relationships. As such, if we understand the component dynamics at a given scale then prediction of their influences will be feasible, allowing for appropriate simulation of their response to shifts in system properties. While theorizing and experimentation are the most appropriate means of elucidating biological truth in ecological dynamics; simulation, especially for microbial communities, represents a new frontier for designing in silico experiments to test fundamental hypotheses. These can, by definition then be tested through observation, experimental manipulation and theory.

Emergent properties are usually defined as the output resulting from an interacting set of variables that have a property, which emerges from their interactions. In microbial ecosystems, emergent properties can be gas fluxes or some host response to changes in microbial assemblage structure, functional potential, transcription, metabolic activity and the physicochemical context of the medium in which those microorganisms reside. One ecosystem emergent property that microbes play a role in is the production of methane. The mechanics of this activity are understood at a rudimentary level, yet the myriad interactions that result in the production of a unit volume of methane from a unit volume of ecosystem remain poorly characterized and very hard to capture. Why is this important? Well, if we can capture the components and their dynamics that lead to the production of methane it is likely possible that we could quantitatively predict, using a given set of state variables, the production of methane for a given ecosystem. In this case, our emergent property is the potential for methane flux to be predictable based on knowledge of the genotypic components, redox chemistry conditions, and

physicochemical variables. However, such multi-scale interactions are rarely amenable to simplified predictive understanding, because small fluctuations in either measured or unmeasured variables can lead to significant shifts in predicted outcome. Using biogeographic field data and longitudinal microcosm studies it is possible to create complimentary streams of evidence that can be used to identify multi-scale interactions, both correlative and putatively causative, that lead to a mechanistic emergent property.

Creating a model of the components of an ecosystem that support an emergent property is not easy to do. It is necessary, for example, to understand the physicochemical dynamics that influence the geochemistry for a given flux. Additionally, you need to capture, characterize and interpret the biological components, define their potential interactions, and determine how these interactions and metabolism influence a flux. However, once you have defined a suite of variables and captured enough information, at multiple scales, about their correlative relationships and how these relationships correlate with changes in a gas flux, it should be possible to generate a prediction just by inputting the values for the system variables you want to predict. A model of such principles across spatial or temporal scales should not ignore fine scale biogeography or biogeochemistry, for variation in these 'elements' of the system will provide significant feedback on larger scale processes.

Even if we can capture and describe in a model the dynamics of a system at a particular scale, and appropriately predict the influence of changes in this scale on larger scales, we still have a dynamic range problem. Dynamic properties below the current scale of simulation may not have been captured, which could lead to inaccurate predictions, as these sub-simulation level process dynamics could influence predicted dynamics. This limits the acceptance of extrapolated understanding for biological phenomena. For example, while it is possible to capture the genotypic diversity for methanogens, methanotrophs and anaerobic respiratory microorganisms across an ecosystem, and generate predictions of how the diversity of these genotypes influences methane flux at the scale of a hectare of land, this still does not capture processes occurring within each organismal unit. In this scenario these sub-simulation components could include incomplete gene-complement annotations (so that we don't know what the organism is fully capable of doing or responding to), epigenetic modifications within a population of cells affecting transcriptional regulation, and micron-scale local environment characteristics influencing cellular physiological responses (e.g. the same bacterium in a gram of soil could be located in one of many different micro-environments with profound influences on it's physiological properties). These uncertainties mean that for any simulation it is essential to thoroughly understand the assumptions and accept those assumptions as limitations in our understanding of the accuracy of extrapolated predictions.

Non-linear dynamics are ubiquitous in nature, and this ubiquity makes linear correlation analyses difficult to justify. The switching relationships between variables in microbial ecosystem dynamics lead to ephemeral correlations that result in inappropriate interpretation of relationships. Applying techniques such as convergent cross mapping (Sugihara et al., 2012) to time series data may overcome this problem by examining the

extent to which temporal components of a given variable can be used to predict the state of another variable. This is achieved by creating a relationship manifold, against which the longitudinal variance of multiple variables is mapped, in this way intrinsic relationships between multiple variables can be established. This technique can distil probable causation from correlation by cross mapping estimates through convergence.

Ideally, we ultimately desire to develop models that extend beyond a statistical interpretation of biological observations in microbial communities to a mechanistic understanding of the interactions and species attributes that truly govern microbial community dynamics. Metabolic models of complex microbial communities are already here in many forms, including compartmentalized models of simple defined communities (Freilich et al., 2011), biogeochemical models with simplified community dynamics (Zhuang et al., 2011), and ensemble models from metagenome data (Larsen et al., 2011). We will increasingly move towards modeling more complex communities that include species that are currently unculturable. These species' genomes will be obtained either by single cell genome sequencing or assembly of metagenomic reads. We will produce increasingly high quality models of the species comprising a community, as methods to produce models from genomic data (Henry et al., 2010) improve; and we will merge those species models into full community models using both dynamic spatial simulation techniques and flux balance techniques. These models will elucidate how species in a community cross-feed and how they carve out and defend their ecological niche through a combination of unique capacity, signaling, attack, and defense. Models will be refined through the integration of metatranscriptomic and metabolomic data, which will confirm and improve predictions of active pathways and metabolites.

A major challenge in producing these mechanistic models of microbial communities is our lack of a mechanistic understanding for many pathways that drive the production and utilization of essential nutrients, signal molecules, or inhibitors that mediate community dynamics. The biochemistry of many of these pathways has yet to be worked out, much less having biochemical transformations mapped to protein sequences. The time is ripe for the application of cheminformatics approaches (Hatzimanikatis et al., 2005) combined with high-throughput functional assays to rapidly fill these knowledge gaps that preclude mechanistic modeling of many phenomena that are important for microbial community dynamics at this time. These cheminformatic approaches automatically generate candidate pathways for production or degradation of a compound of interest; we can then map those candidate pathways to potential enzymes selected from known genome sequences by drawing on theories for prediction of enzyme promiscuity (Khersonsky and Tawfik, 2010). Our community metabolic models may also serve as test-beds for identifying the pathway candidates that best fit with observed community behavior and metatranscriptomic and metabolomic data. This combination of approaches will greatly enhance our mechanistic understanding of microbial community dynamics, permitting longer-range prediction of behavior and enabling the engineering and redesign of communities.

The time has come where we can, at certain levels of system complexity, capture the dynamics of enough variables to predict the systems emergent properties. Capturing these

and modeling them in silico to produce effective predictions is not that distant a dream. We predict that within the next 5 years we will start to see specific models with defined components and system properties that can be taught as a neural net, learning to respond to environmental stimuli. By focusing on the microbial components of these prospective models, it is likely possible that we could forecast how emergent properties will change under given climate change scenarios, or pollution events, or disease outbreaks. If we build the infrastructure, and design our experiments carefully to gather the appropriate data, this is not unachievable.

Acknowledgements

This work was supported in part by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357.

References

Freilich, S., Zarecki, R., Eilam, O., Segal, E.S., Henry, C.S., Kupiec, M., et al. (2011) Competitive and cooperative metabolic interactions in bacterial communities. *Nat. Commun.* **2**: 589.

Hatzimanikatis, V., Li, C., Ionita, J.A., Henry, C.S., Jankowski, M.D., and Broadbelt, L.J. (2005) Exploring the diversity of complex metabolic networks. *Bioinforma. Oxf. Engl.* **21**: 1603–1609.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**: 977–982.

Khersonsky, O. and Tawfik, D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**: 471–505.

Larsen, P.E., Collart, F.R., Field, D., Meyer, F., Keegan, K.P., Henry, C.S., et al. (2011) Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb. Inform. Exp.* **1**: 4.

Sugihara, G., May, R., Ye, H., Hsieh, C. -h., Deyle, E., Fogarty, M., and Munch, S. (2012) Detecting Causality in Complex Ecosystems. *Science* **338**: 496–500.

Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D.R. (2011) Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *ISME J.* **5**: 305–316.